自然言語処理 114－14
(1996. 7. 19)

確率文脈自由文法が持つ解析木の生成数抑制能力に関する検証

渥美 清隆　　増山 繁

豊橋技術科学大学 知識情報工学系

〒441, 愛知県豊橋市天伯町字雲雀ヶ丘 1-1

あらまし

確率文脈自由文法に基づく k best parser は 1 つの入力文に対して, 多数の解析木が生成されるような曖昧性を制御する能力を備えている. 本論文では, この曖昧性の抑制能力を検証するために, 解析木に付与される確率値の出現割合に関する近似計算と, 日本語コーパスから学習した確率付きの文法を用いた実験を行ない, 比較検討した.

キーワード

確率文脈自由文法, k best parser, 曖昧性, 解析木

A Research on Ability of Reducing the Amount of Parsing Trees Generated by a Stochastic Context-Free Grammar

Kiyotaka ATSUMI　　Shigeru MASUYAMA

Dept. of Knowlege-based Information Engineering, Toyohashi Univ. of Tech.

1-1, Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi-ken, 441 Japan.

Abstract

This paper verifies the ability of k best parser for stochastic context-free generated to reduce the ambiguity of the results. Here, ambiguous results means to generate a number of parsing trees for one sentence. We take two method for this verification. One is an approximation for the distribution of probability assigned to a parsing tree, the other is an experiment using a stochastic grammar generated by learning system from Japanese corpus. And then we compare the results by these two methods.

Key Word

stochastic context-free grammar, k best parser, ambiguous, parsing tree

1 はじめに

自然言語の分析のために, 文脈自由文法に基づいた構文解析を利用することが多い. しかし, 文脈自由文法には曖昧性が含まれることが多く, この場合 1 つの入力列に対して, 複数の解析結果を構文解析が出力することになる. 解析結果の数があまり多くならなければ, 意味解析などに選択を委ねることも可能であるが, 非常に多くの解析結果を出力することになった場合には, 意味解析だけで選択することは難しく, それ以外の簡便な選択方法が必要になる.

そこで, 各解析結果に対して確率値を付与し, その確率値の大きさによって解析結果を絞り込むことが考えられる. これは確率文脈自由文法 [1] によって実現できるが, 解析結果をどの程度絞り込むことができるのかを議論した論文は我々の知る限りでは従来にない.

本論文では, 解析木に付与される確率についていくつかの仮定を置き, 解析木の確率値の出現確率の密度関数と重み付きの密度関数を計算することにより, 解析結果をどの程度絞り込むことができるのかを示す近似計算について議論し, さらに文法の自動学習を利用してコーパスから日本語確率文脈自由文法を獲得して, それを用いた実験との比較検討を行う.

2 確率文脈自由文法 [1] の定義

ここでは, 確率文脈自由文法 $SCFG$ (Stochastic Context-Free Grammar) と解析木への確率の付与の仕方について定義する.

2.1 確率文脈自由文法の定義

確率文脈自由文法は $SCFG = (N, \Sigma, S, P, p)$ の 5 つ組で表現され, それぞれ以下のように定義する.

- N は非終端記号の集合,
- Σ は終端記号の集合,
- S は出発記号,
- P は導出規則の集合,
- p は $P \to [0, 1]$ の写像,

導出規則は

$$A \underset{p(A \to \beta)}{\to} \beta,$$

$$A \in N,\ \beta \in (N \cup \Sigma)^*$$

で表現される. 各導出規則に確率値 p が割り当てられていることを除けば, 通常の一般文脈自由文法と全く同じである. $p(A \to \beta)$ は導出規則 $A \to \beta$ を使うときの非終端記号 A に対する条件付き確率である. A から導出される規則が k 個あり, それぞれの導出規則を $A \to \beta_i,\ 1 \leq i \leq k$ とするとき, $\sum_{i=1}^{k} p(A \to \beta_i)$ は 1 で正規化されている.

2.2 確率文脈自由文法から得られる解析木の確率の定義

確率文脈自由文法を用いた構文解析は一般の文脈自由文法の構文解析と何等変わることはない. しかし, 構文解析を行った後, それぞれの解析木に確率値を付与するために再計算を行う. 各解析木 T はいくつかの導出規則 $P_1, \cdots, P_i$ から成り立っており, それぞれの導出規則には確率値 $p(P_1), \cdots, p(P_i)$ が付与されている. これらの確率値から, 以下のような計算をする.

$$p(T) = \prod_{j=1}^{i} p(P_j)$$

この計算により各解析木の確率値が与えられる.

3 解析木に付与される確率値の近似

我々が興味を持っている情報は, 構文解析から出力される多数の解析木のうち, 確率的にみて意味のある解析木がどのくらい存在しているかである.

導出規則は解析木の節点に付与される非終端記号に従う. この導出規則の使用頻度や使用位置は入力列に応じた複雑な振る舞いをするため, 具体的な入力列が定まらない限りどのような導出規則が使用されるかは特定できない. このため, 構文解析が出力する解析木に付与される確率値の情報, 例えば a から 1 までの範囲の確率値を持つ解析木が, 各入力列の平均でいくつ存在するのか等の情報を得ることを困難にしている.

解析木に付与される確率値の振る舞いを複雑にしている理由を整理すると,

1. 一般の文脈自由文法では解析木に使用される導出規則の数が特定できない,
2. 導出規則の使用頻度や位置は入力列に依存する,
3. 導出規則に付与されている確率値の分布情報が離散的である,

ことである. これらの問題を解決するために以下の3つの仮定を置く.

1. 文法はChomsky標準形で書かれている.
2. それぞれの導出規則が使用される頻度は一様であり, 位置は互いに独立である.
3. 導出規則に付与されている確率値の出現確率を示す確率変数は微分可能な密度関数で表現できる.

1つ目の仮定を置く目的は, 1つの入力列から得られる解析木の使用する導出規則の数を固定するためである. Chomsky標準形で書かれた文法を利用して構文解析を行うと, 入力列の長さがnのとき, 解析木に使われる導出規則の数mは$m = 2n - 1$個に固定することができる.

2つ目の仮定を置く目的は, 導出規則間の従属関係による複雑な振る舞いをする解析木に付与される確率値の計算を, 単純な振る舞いに変換するためである. ここで単純化計算法の概略を説明する. 導出規則に付与されている確率値をボールに1つづつ書き写し, それを大きな袋に入れることを想定する. もし, 入力列の長さがnであるならば$m = 2n - 1$個のボールを復元抽出で袋から取り出し, ボールに記録されている数字を書き取る. この数字の全ての積がその入力列から得られるある解析木に付与される確率値であるとする. この試行を何回か繰り返したときに得られる密度を解析木に付与される確率値の出現確率の密度とする.

3つ目の仮定を置く目的は, 離散的な振る舞いから連続的な振る舞いに換えることにより, 微積分などの手段を利用可能にすることである.

3.1 評価方法の定義

導出規則に割り当てられた確率値の出現確率が確率変数X_1に従うとする. この確率変数の密度関数が$f_1(x)$で表現できるとき, m個の導出規則を使った解析木に付与される確率値の出現確率は確率変数X_mに従う. この確率変数の密度関数を

$$f_m(x) = \int_x^1 f_1(\frac{x}{y}) f_{m-1}(y) \frac{1}{|y|} dy \qquad (1)$$

と定義する. これはm個の独立した確率変数から得られる確率の積と同一である[2]. また, $xf_m(x)$が重み付きの密度関数であるとするとき,

$$F_{m,a}(x) = \int_a^1 x f_m(x) dx \qquad (2)$$

はaから1までの確率値が付与された解析木の重みの合計を表わす.

(2)式においてaから1までの解析木の重みの合計がrに達するようなaを求めれば, 解析木に付与される確率の大きい順に解析木を選択するとき, いくつまで選択すればよいのかの目安にすることができる. このaを求める式として以下を定義する.

$$\max_a \left\{ \frac{F_{m,a}(x)}{F_{m,0}(x)} \geq r \right\} \qquad (3)$$

そして, (3)式で求めたaから, 全体の解析木の数と比較してどのくらいの割合で解析木を取り出せばよいかは, 次の式に従う.

$$RNT_{m,a}(x) = \int_a^1 f_m(x) dx \qquad (4)$$

この(4)式から得られる数がmと共にどのように変化するかを考察する. 以下の節では導出規則に付与される確率値の出現確率の密度関数について具体的な連続関数を想定して考察する.

3.2 分布関数が一様分布の場合

導出規則に付与される確率値の出現確率の密度関数が$f_1(x) = 1, 0 \leq x \leq 1$に従うとき, つまり確率値が$0.1,\ 0.2, \cdots, 1.0$のように等間隔に存在する場合を考える. このとき$f_2(x)$は(1)式から

$$f_2(x) = \int_x^1 f_1(\frac{x}{y}) f_1(y) \frac{1}{|y|} dy = -\log x$$

が得られる. この式が持つaから1までの重み付き密度関数の積分値は

$$F_{2,a}(x) = \int_a^1 x f_2(x) dx = \frac{1}{4} + \frac{a^2}{2}(\log a - \frac{1}{2})$$

となる．$a = 0$ のとき，この式の値は 1/4 になるので，$r = 0.8$ としてこれを (3) 式に代入すれば，

$$\max_a \left\{ \frac{\frac{1}{4} + \frac{a^2}{2}(\log a - \frac{1}{2})}{\frac{1}{4}} \geq 0.8 \right\}$$

となる．この式から a は約 0.22 と求めることができる．そしてこの a の値を (4) 式に代入すると，

$$RNT_{2,0.22}(x) = -\int_{0.22}^1 \log x dx = 0.447$$

を得る．つまり，構文解析から出力される解析木のうち，解析木に付与された確率が大きい順に，解析木の総数の 44.7% を選択すればよいことになる．

さて，この議論が実際の導出規則が持つ確率値の振る舞いと一致するかどうかの簡単な検証を行う．まず，導出規則に付与された確率値は $\hat{f}_1 = \{0.1,\ 0.2, \cdots, 1.0\}$，$|\hat{f}_1| = 10$ であったとする．導出規則を独立に 2 つ選択し，それぞれが持つ確率値の積を求めたとき，$\hat{f}_2 = \{0.01,\ 0.02,\ 0.02,\ 0.03, \cdots, 0.81,\ 0.9,\ 0.9,\ 1.0\}$，$|\hat{f}_2| = 100$ を得る．この $\hat{f}_2$ の集合から大きい順に 49 個を選択すると全体の重みの 80.3% に達する．(4) 式に対応する計算では $49/100 = 0.49$ となるので，連続関数上での議論とほぼ一致するものと考える．連続関数上の議論と離散関数上での議論の違いは最初の導出規則に付与される確率値の密度関数の微妙な差から生まれていることが考えられる．

次に入力列の長さを制限しない一般の場合について検討する．$f_2(x)$，$f_3(x), \cdots$ と順に求めていくと，

$$f_m(x) = \frac{(-1)^{m-1}}{(m-1)!} \log^{m-1} x$$

を得ることができる．この式を (2) 式に代入すると，

$$\begin{aligned} F_{m,a}(x) &= \int_a^1 x f_m(x) dx \\ &= \frac{(-1)^{m-1}}{(m-1)!} \left\{ \left[\frac{x^2 \log^{m-1} x}{2} \right]_a^1 \right. \\ &\quad \left. - \frac{m-1}{2} \int_a^1 x \log m - 2x dx \right\} \\ &= \frac{(-1)^{m-1}}{(m-1)!} \left[\frac{x^2 \log^{m-1} x}{2} \right. \\ &\quad - \frac{m-1}{2} \left(\frac{x^2 \log^{m-2} x}{2} \right. \\ &\quad - \frac{m-2}{2} \left(\frac{x^2 \log^{m-3} x}{2} \right. \\ &\quad \vdots \\ &\quad - \frac{2}{2} \left(\frac{x^2 \log x}{2} \right. \\ &\quad \left. \left. \left. - \frac{1}{2} \left(\frac{x^2}{2} \right) \cdots \right) \right]_a^1 \end{aligned} \tag{5}$$

となる．また，(4) 式に代入すると，

$$\begin{aligned} RNT_{m,a}(x) &= \int_a^1 f_m(x) dx \\ &= \left[\frac{(-1)^{m-1}}{(m-1)!} x \log^{m-1} x \right. \\ &\quad + \frac{(-1)^{m-2}}{(m-2)!} x \log^{m-2} x \\ &\quad \vdots \\ &\quad + \frac{(-1)^1}{1!} x \log x \\ &\quad \left. + \frac{(-1)^0}{0!} x \right]_a^1 \end{aligned} \tag{6}$$

となる．本来なら (5) 式を (2) 式に代入し，一様分布における a を求める n の関数式を作り，その関数式を (4) 式に代入するべきであるが，単純な式にまとめることが難しいため，コンピュータ上で $m = 1 \sim 100$ までについて a を計算し，その結果の一部を表 1にまとめた．また，ここで計算されたそれぞれの a の値を (6) 式に代入した結果の一部を表 2にまとめた．表中の r は確率の大きい順に解析木を選択した場合の重みと解析木の重み全体に対する比率である．

表 1, 2から入力列長 n の増加量に対して，どちらも指数的に減少していることが分かる．特に，解析木の選択割合の指数的な減少は，入力列長に応じて指数的に増加する解析木を抑制する効果が期待できる．

3.3 分布関数が一様分布以外の場合

導出規則に付与される確率値の出現確率の密度関数の一般的な性質として，密度関数の平均が 0.5 を超えることがほとんど有り得ない，ということが挙げられる．

表 1: 一様分布の場合の各入力列長 n(導出規則の数 m)に対する a の値

$n(m)$	$r=0.2$	$r=0.5$	$r=0.8$
2(3)	0.464	0.262	0.118
5(9)	0.0402	0.0131	0.00338
10(19)	0.000484	8.84e-5	1.28e-5
15(29)	5.03e-6	5.96e-7	5.57e-8
20(39)	4.88e-8	4.01e-9	2.61e-10
25(49)	4.52e-10	2.70e-11	1.28e-12
30(59)	4.06e-12	1.82e-13	6.45e-15
35(69)	3.57e-14	1.23e-15	3.34e-17
40(79)	3.07e-16	8.27e-18	1.76e-19
45(89)	2.60e-18	5.57e-20	9.42e-22
50(99)	2.18e-20	3.76e-22	5.11e-24

表 2: 一様分布の場合の各入力列長 n(導出規則の数 m)に対する解析木の選択割合

$n(m)$	$r=0.2$	$r=0.5$	$r=0.8$
2(3)	0.0429	0.151	0.361
5(9)	0.00587	0.0331	0.122
10(19)	0.000375	0.00357	0.0219
15(29)	3.02e-5	0.000431	0.00395
20(39)	2.71e-6	5.48e-5	0.000705
25(49)	2.61e-7	7.16e-6	0.000124
30(59)	2.63e-8	9.52e-7	2.18e-5
35(69)	2.75e-9	1.28e-7	3.78e-6
40(79)	2.94e-10	1.74e-8	6.52e-7
45(89)	3.22e-11	2.39e-9	1.12e-7
50(99)	3.59e-12	3.29e-10	1.90e-8

なぜならば, 確率値 1 を持つ導出規則を除いた導出規則に付与される確率値の平均は必ず 0.5 以下であり, 確率値 1 を持つ導出規則は通常ごく僅かであると考えられるからである. また, 導出規則の確率値の出現確率の密度関数の平均が μ であるとき, 解析木の確率値の出現確率の密度関数の平均は入力列長が n であるならば μ^{2n-1} になる. そして, 解析木の確率値の出現確率の密度関数を導出規則の組み合わせとして考えるとき, 組み合わせたときの調整パラメータとして働く $1/|y|$ の項が log の多項式的に変化するため, n がある程度大きくなると log の多項式が他の項に比べて支配的になる. このため, 導出規則に付与される確率値の出現確率の密度関数がどのような関数であろうと, 一様分布の場合とあまり変わらず, n が大きくなるに従って, 解析木の選択割合は指数的に減少することが期待できる.

4 日本語コーパスを用いた仮定の妥当性の検証

我々は確率文脈自由文法が生成する確率付き解析木の確率値の振る舞いについて, 数値計算による近似計算を行うための 3 つの仮定を置いた. この節では 3節で議論した近似計算が実際の確率文脈自由文法の振る舞いをどの程度表現しているのかについて実験を通じて議論する.

4.1 日本語確率文脈自由文法の獲得

日本語のための確率文脈自由文法は一般に公開されていない. そこで日本語文法をコーパスから直接学習する方法について考える. 文法の自動獲得については [4] などで議論されている. 本研究で用いたこの文法の獲得方法の概要は次のような手順である.

1. すべての入力文に対して解析木をランダムに割り当てる. つまり解析木の形もランダムであり, 解析木の各節点に割り当てられる非終端記号もランダムである. この状態を最初の暫定解とする.

2. 暫定解から導出規則を抜き出し, エントロピー H を以下のように計算する.

$$H = -\sum_{i,j} \hat{p}(N_i \to T_j) \log_2 p(N_i \to T_j)$$
$$-\sum_{i,j,k} \hat{p}(N_i \to N_j N_k) \log_2 p(N_i \to N_j N_k)$$

を計算する. このとき $\hat{p}$ は導出規則全体に対する使用頻度の (つまり条件付きではない) 確率である. また N と T はそれぞれ暫定解中に現れた非終端記号, 終端記号とする.

3. ある入力列に対して別のランダムな解析木を割り当て, その部分を交換した場合の H を計算し, H が下がる場合には交換した解析木を採用し, そうでない場合は解析木を元に戻して新しい暫定解とする.

4. 2 と 3 を規定回数繰り返す.

この方法は局所探索 (local search)[3] と呼ばれる方法であり, この方法で得られる文法は, 学習が局所解に陥らずに進めば, 曖昧性がもっとも少ない文法を得ることが可能である. また各導出規則にはすでに確率が付与されていることになるので, 本研究の実験でもその確率をそのまま採用している.

4.2 学習結果の文法の性質

先の文法学習法を利用し, 実際のコーパスから文法の学習を試みた. コーパスは日経新聞 90 年度版 CD-ROM の社説から 100 文を選択し, JUMAN により形態素解析を行い品詞の列に変換した結果を入力文とした. 入力文のサイズは実際に作成したプログラムと計算機資源の兼ね合いで決定した.

初期のエントロピー H は 8.41 であり, 学習を 200 万回繰り返した後ではエントロピー H は 7.45 になった. エントロピーの下がり具合があまりよくないのは入力文数が少ないためであると考えられる. 最終的に出力された導出規則の数は Chomsky 標準形で 440 個, 付与された確率値の平均は 0.0455 である.

4.3 実際に生成された解析木に付与された確率値の性質と近似計算との比較

品詞列の長さが 5 および 9 の各入力文に対して, 自動学習から得られた文法に基づき構文解析を行ない, 全ての確率付解析木を出力させた. 結果は表 3, 4, 5の通りである.

表 3: 構文解析の結果

	長さ	全解析木数
文 1	5	62689
文 2	9	13094

表 4: 各文に対する重みの割合 r に達っする a の値

	$r = 0.2$	$r = 0.5$	$r = 0.8$
文 1	1.47e-9	5.41e-10	1.70e-10
文 2	1.38e-20	6.01e-20	2.29e-21

表 5: 各文に対する重みの割合 r に達する解析木数の割合

	$r = 0.2$	$r = 0.5$	$r = 0.8$
文 1	0.0141	0.0766	0.252
文 2	0.0204	0.0977	0.278

解析木全体からどのくらいの割合で確率の大きい順に解析木を取るのかについて表5を見てみると, 文 1 から文 2 に文の長さが変化すると, その長さの変化の割合よりはやや少ない増加をしている. ところが, 表 2からでは, いづれも指数的減少を示している.

この違いは, 自動学習した文法の曖昧性によるところが大きいと考えられる. 事実, エントロピーの計算上, $H = 7.45$ はある節点の非終端記号から平均 7.45 個の導出規則の選択の余地がある. 文献 [4] では, この H が約 2 程度にまで減少していることを考えれば, 自動学習についてもう少し工夫しなければならない.

5 まとめ

本論文では, 確率文脈自由文法に基づく構文解析を行なった結果の解析木に付与される確率値の振る舞いについて, 近似計算と文法の自動獲得による実験の 2 つの側面から検討した. 近似計算においては3節の 3 つの仮定の下で, 以下のことを明らかにした.

1. 確率値が非常に小さな解析木が大多数を占め, 確率的に意味のある解析木は非常に僅かである.

2. 入力列が長くなると, 全体の解析木の数に比べて確率的に意味のある解析木は指数的に少なくなる.

3. 一様分布だけでなく, 導出規則に付与される確率値の密度がいかなる関数であっても, 上記2つの性質を期待することができる.

ところが, 自動獲得した文法による実験では, 近似計算における2番目と3番目の事実と反する結果となっている. 近似計算では, 文法がどの程度の曖昧性を持っているのかについて全く考慮していなかったことと, 自動学習から得られた文法には非常に多くの曖昧性が含まれていたこととの相乗効果により, このような結果になったと考えられる.

これらの問題点から今後の課題として,

1. 近似計算において, 文法の持つエントロピーを考慮,

2. 近似計算における解析木に付与されている確率値の合計に対して, 解析木に付与されている確率値の大きい順に解析木を選択するときの割合, つまり $r = 0.2,\ 0.5,\ 0.8$ という数字の現実的妥当性の検討,

3. 文法に付与されている確率値の精度の向上,

が挙げられる. これらの改善を行なった上でさらに, 確率文脈自由文法が持つエントロピーがどの程度であれば, 確率的に意味のある解析木の現実的な絞り込みができるのかについて検討する.

参考文献

[1] 中川 聖一: "確率モデルによる音声認識," 信学会 (1988).

[2] A.M.Mood, F.A.Graybill and D.C.Boes: "Introduction to the Theory of Statistics," *McGraw-Hill* (1974). 和訳 大石: 統計学入門 (上) マグロウヒル好学社(1978).

[3] 茨木 俊秀: "アルゴリズムとデータ構造," 昭晃堂 (1989).

[4] 横田 和章 他: "コーパスに基づく日本語文法の自動獲得," 言語処理学会第2回年次大会論文集, pp.169-172 (1996).